Attention Transfer Is Not Universally Effective for Vision Transformers

Huaiyuan Qin [1] **Muli Yang** [1] **Gabriel James Goenawan** [2,†]
Peng Hu [3] **Chen Gong** [4] **Xi Peng** [3] **Hongyuan Zhu** [1,✉]

[1]Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore
[2]Intrinsic, Google [3]School of Artificial Intelligence, Sichuan University
[4]School of Automation and Intelligent Sensing, Shanghai Jiao Tong University

{qinhy, yangml, zhuh}@a-star.edu.sg, gabrieljames@google.com,
{penghu.ml, pengx.gm}@gmail.com, chen.gong@sjtu.edu.cn

🔗 **Project Page**

Abstract

A recent work shows that Attention Transfer, which transfers only the attention patterns from a pre-trained teacher Vision Transformer (ViT) to a randomly initialized standard student ViT, is sufficient to recover the full benefit of the teacher's pre-trained weights. We revisit this finding on a comprehensive benchmark of 20 teachers from 11 well-known ViT families and reveal that Attention Transfer is not universally effective. While 7 families transfer successfully, 4 consistently fail, falling up to 5.1% below the from-scratch no-transfer baseline. Further results demonstrate that this failure is family-consistent across model sizes, and persists under extended training durations, different transfer datasets, and out-of-distribution evaluations. Controlled analyses then consistently localize the problem to the attention-routing channel, indicating that the key issue is not whether the student can match the teacher's attention patterns, but whether the matched patterns remain functional for the student. Crucially, we identify architectural mismatch between the pre-trained teacher and the standard student as the primary mechanism. By adding only the teacher's native architectural components to the student in a randomly initialized state, we completely reverse the failure for all 4 families. Notably, these components alone do not improve from-scratch training, confirming that they specifically unlock the usability of the teacher's attention. We further show that this failure is not explained by the inadequate choice of transfer loss, differences in pre-training recipes, or the intrinsic quality of attention maps. Our findings refine the prevailing understanding of attention in ViT representations: attention is sufficient *only* when the student architecture matches the teacher.

1 Introduction

Pre-trained Vision Transformers (ViTs) [1] have become the standard foundation for modern visual representation learning, but obtaining such weights typically requires massive computation and large curated datasets [2–4]. This makes the efficient reuse of pre-trained representations a central objective in computer vision. In practice, transferring a pre-trained ViT usually means initializing a downstream model with the full set of teacher weights. While effective, it remains fundamentally unclear which specific components of a pre-trained ViT actually encode its transferable knowledge [5–7].

† This paper was completely accomplished when GJG was at A*STAR.
✉ Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

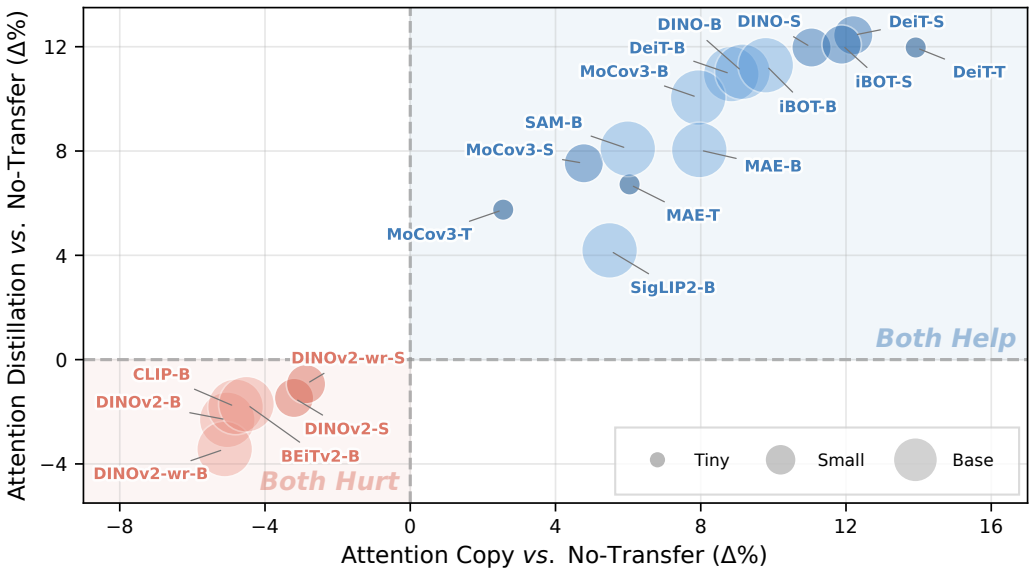

Figure 1: **Attention Transfer is not universally effective.** We begin by evaluating Attention Transfer [8] on the broader model zoo of 20 pre-trained teachers from 11 well-known ViT families under the default ImageNet-1K setup. Each point represents one set of pre-trained weights; the axes report Δ Top-1 accuracy (%) relative to the No-Transfer baseline under two transfer methods: *Attention Copy* (x-axis) and *Attention Distillation* (y-axis). Marker size denotes the model size (Tiny/Small/Base). 7 families consistently succeed (in the blue top-right region), while 4 consistently fail (in the red bottom-left region) under *both* transfer methods.

One recent work [8] offers a striking answer to this question. It shows that *Attention Transfer*, which transfers only the teacher's attention patterns to a randomly initialized standard student ViT through attention copy or distillation, is sufficient to recover the full benefit of the teacher's pre-trained weights. This result suggests that the essential transferable structure of a pre-trained ViT lies in its inter-token routing patterns, while the remaining parameters can be learned from scratch once this attention behavior is imposed. A natural question is whether this finding holds broadly across the increasingly diverse landscape of pre-trained ViTs, and more fundamentally, whether attention patterns are truly a portable representation of pre-trained knowledge across architectures.

In this work, we show that this ***does not*** hold universally. By revisiting Attention Transfer on a substantially broader and more modern model zoo of 20 teachers from 11 well-known ViT families [2, 4, 9–17], we uncover a hidden validity boundary that was invisible under the evaluations in [8]: its effectiveness is highly family-dependent. Under the standard teacher-student transfer protocol used in [8], Attention Transfer consistently succeeds for 7 families, but consistently fails for 4. For these failure families, transferring attention yields strictly worse downstream performance than the from-scratch no-transfer baseline, with the worst case falling 5.1% below scratch. Further results show that this failure is family-consistent across model sizes, and persists under extended training durations, different transfer datasets, and out-of-distribution evaluations.

To understand why these failures arise, we conduct a set of controlled analyses, including component-level, layer-wise, and QKV decomposition ablations, which consistently localize the problem to the attention-routing channel, the part that controls inter-token information flow in ViTs, under a standard student architecture. This shows that the key issue is not simply whether the student can match the teacher's attention patterns, but whether the matched patterns remain functional in the student's attention routing. This naturally raises a sharper question: what determines this functional usability across teacher families?

Crucially, we identify *architectural mismatch* between the pre-trained teacher and the standard student as the missing factor. The standard student ViT used in [8] omits architectural components that are native to those failure families, such as `LayerScale` in DINOv2 [2] and DINOv2-wr[1] [17], `PreLayerNorm` in CLIP [4], and `LayerScale` together with `RelativePositionBias` in BEiTv2 [14]. Moreover, by simply adding the teacher's native architectural components to the student in a *randomly*

[1]We term the family of DINOv2-with-registers [17] as DINOv2-wr in the following for simplicity.

initialized state, we completely reverse this failure for all 4 families. Importantly, training this native-architecture student entirely from scratch yields worse performance than a standard from-scratch student, confirming that these components do not artificially boost performance on their own. Instead, they specifically make the teacher's transferred attention usable, showing that attention is sufficient to recover the full benefit of the teacher's pre-trained weights *only* when the student architecture matches the teacher.

We then ask whether simpler explanations could account for these failures. We investigate three natural alternative explanations: the choice of inadequate transfer loss, differences in pre-training recipes, and the intrinsic quality of attention maps. First, through empirical analysis, we find that different losses with sufficient magnitude can induce similar supervision signals for Attention Transfer. However, for the failure families, increasing the magnitude of any loss further hurts performance, showing that the loss itself does not explain the failure. Second, we analyze several hypotheses based on common properties in pre-training recipes of the failure families, including different training signals, different training data sources, and other known special attributes, but find none that explain this family-level boundary. Lastly, we measure the quality of the teachers' attention maps by evaluating their object localization accuracy, finding that the quality ranking does not predict the observed failure, and that the native-architecture rescue, which reverses the failure, does not consistently change this quality.

In summary, our primary contributions are as follows:

1. **Failure Discovery.** We provide the first comprehensive Attention Transfer [8] benchmark across 20 teachers and 11 well-known ViT families, uncovering a critical validity boundary where 4 major families consistently fail to transfer to a standard student architecture.
2. **Three-Level Localization.** Through systematic component-level, layer-wise, and QKV decomposition ablations, we localize the failure specifically to the attention-routing channel, showing that the key issue is whether the matched patterns remain functional for the student.
3. **Architectural Compatibility as Mechanism & Fix.** We identify architectural mismatch as the primary mechanism behind Attention Transfer failure. Adding the teacher's native components to the student in a randomly initialized state completely reverses the failure for all 4 families, while these components alone hurt from-scratch training.
4. **Elimination of Alternative Explanations.** We analytically and empirically rule out potential inadequate choice of transfer loss, pre-training recipe differences, and the intrinsic quality of attention maps as explanations for the observed failure boundary, strengthening the finding that architectural mismatch is the primary explanation of Attention Transfer failure.

2 Preliminaries

2.1 Attention Transfer

Following [8], we study *Attention Transfer*, where a randomly initialized student Vision Transformer (ViT) [1] is trained to inherit the full benefit of a pre-trained teacher without loading the teacher's full pre-trained weights. Specifically, for a ViT block ℓ, given Q, K, and V as the corresponding queries, keys, and values, the attention function is defined as[2]:

$$f_{\text{attn}}^{(\ell)} = \underbrace{\text{softmax}\left(Q^{(\ell)}K^{(\ell)\top}\right)}_{\text{attention map}} V^{(\ell)}, \tag{1}$$

where the softmax function is computed per query for the *attention map*.

We study two forms of Attention Transfer from a pre-trained teacher ViT to a standard student[3] of the same size: ***Attention Copy***, which directly copies the teacher's attention projection weights and keeps them frozen; and ***Attention Distillation***, which trains the student to match the teacher's attention patterns through an auxiliary transfer objective. Given $Q_sK_s^\top$ and $Q_tK_t^\top$ from the student and teacher, respectively, the transfer loss at block ℓ is defined as:

$$\mathcal{L}_{\text{transfer}}^{(\ell)} = \mathcal{L}\left[\text{softmax}\left(Q_s^{(\ell)}K_s^{(\ell)\top}\right), \text{softmax}\left(Q_t^{(\ell)}K_t^{(\ell)\top}\right)\right], \tag{2}$$

where $\mathcal{L}$ denotes Cross-Entropy loss by default in [8].

[2]We omit the scaling factor $1/\sqrt{d_k}$ for simplicity.

[3]The standard student denotes the vanilla ViT architecture of [1, 9], which is kept fixed regardless of the teacher's architecture; only the teacher's attention is copied or distilled into it.

2.2 Experimental Setup

Model Zoo. We revisit Attention Transfer on a comprehensive model zoo which consists of 20 pre-trained weights from 11 well-known ViT families, including DeiT [9], DINO [10], MoCov3 [11], MAE [12], iBOT [13], SigLIP2 [15], SAM [16], DINOv2 [2], DINOv2-wr [17], CLIP [4], and BEiTv2 [14], spanning three model sizes: Tiny, Small, and Base.

Evaluation Protocol. Unless specifically mentioned, all Attention Transfer experiments are conducted on ImageNet-1K classification [18] with the default protocol in [8] using the standard teacher-student architecture. To carefully isolate the effects of the transferred attention patterns, we maintain a strictly controlled training protocol across all experiments: models are trained for 20 epochs with the AdamW optimizer [19]. Other training hyper-parameters follow the defaults in [8]. All reported results are averaged over 3 random seeds to ensure statistical significance. More details are in Sec. A.

No-Transfer Baseline. As a primary reference, we define the *No-Transfer* (NoT) baseline, where the same standard ViT is randomly initialized and trained from scratch for 20 epochs. Attention Transfer from one given pre-trained weights is considered a "failure" when the final accuracy falls below this baseline, indicating that imposing attention patterns from the teacher is less effective than learning from scratch under the same student architecture.

3 Attention Transfer Is Not Universally Effective

3.1 Main Results

We begin by evaluating the two Attention Transfer methods on the full benchmark of 20 pre-trained teachers from 11 well-known ViT families under the default ImageNet-1K setup. Fig. 1 summarizes the performance of both *Attention Copy* and *Attention Distillation*, measured as Top-1 accuracy gap relative to the No-Transfer baseline. Points in the upper-right quadrant therefore indicate teachers for which both transfer methods are beneficial, whereas points in the lower-left (failure) quadrant indicate teachers for which both transfer methods are harmful.

The results reveal a clear family-dependent split, and this split boundary is consistent across model sizes rather than driven by isolated pre-trained weights. All pre-trained weights maintain a consistent sign under both Attention Transfer methods across all seeds. Specifically, 7 families, including DeiT, DINO, MoCov3, MAE, iBOT, SigLIP2, and SAM, improve over the No-Transfer baseline, with gains reaching 13.9% for DeiT-T under Attention Copy. In contrast, 4 families: DINOv2, DINOv2-wr, CLIP, and BEiTv2, consistently fall into the failure region. The worst case is DINOv2-wr-B, which underperforms the No-Transfer baseline by 5.1% under Attention Copy, showing that imposing the attention patterns of certain teachers can substantially degrade learning under the same student architecture. Notably, all teachers evaluated in [8] belong exclusively to our success group, while the failure families are absent from their evaluations. These results show that Attention Transfer is not universally effective under the standard teacher-student transfer protocol.

3.2 Robustness of the Failure

A natural counter-hypothesis is that the failure families might suffer from slower convergence under the transferred attention patterns, thus extending the training duration or changing the transfer dataset could eventually close or even reverse the performance gap. To rule this out, we further rigorously test the robustness of this observed failure using Attention Distillation on a representative subset from two success (DeiT, DINO) and two failure (DINOv2, DINOv2-wr) families with model size Small.

Extended Training Robustness. We extend the training schedule from 20 to 100 epochs for both the success and failure families. As shown in Fig. 2, the performance gap persists throughout

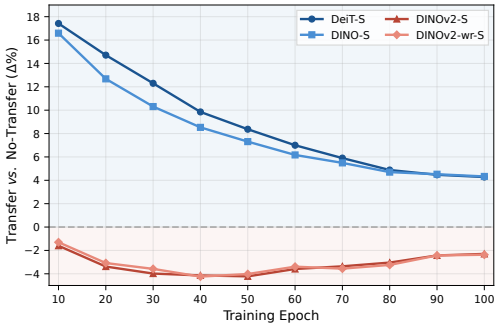

Figure 2: **Extended training robustness.** We extend the Attention Distillation training up to 100 epochs and report the gap relative to No-Transfer every 10 epochs. The same success/failure split remains consistent with the results in Fig. 1.

Table 1: **Dataset robustness.** We evaluate Attention Distillation on iNaturalist [20] with 20-epoch training. Gray row shows the default ImageNet-1K results for reference. Colored deltas indicate gains/drops relative to No-Transfer (NoT) baseline. The success/failure split remains consistent with different datasets.

	NoT	DeiT	DINO	DINOv2	DINOv2-wr
IN-1K	63.1	75.6 +12.4	75.1 +12.0	61.7 -1.5	62.2 -0.9
iNat 2017	26.6	44.7 +18.1	46.0 +19.4	24.4 -2.2	24.5 -2.1
iNat 2018	35.0	50.2 +15.2	51.2 +16.3	33.1 -1.9	33.8 -1.2
iNat 2019	40.1	56.8 +16.7	55.7 +15.5	38.1 -2.1	38.7 -1.4

Table 2: **Out-of-distribution robustness.** We evaluate Attention Distillation on four distribution-shift benchmarks [21–24] with 100-epoch training. Colored deltas indicate gains/drops relative to No-Transfer (NoT) baseline. The success/failure split remains consistent with the in-distribution evaluation.

	NoT	DeiT	DINO	DINOv2	DINOv2-wr
IN-A [22]	9.2	17.6 +8.4	17.2 +8.0	6.0 -3.2	5.8 -3.4
IN-R [21]	35.9	42.3 +6.4	42.8 +6.9	32.8 -3.1	32.2 -3.7
IN-S [23]	22.1	29.7 +7.6	29.4 +7.3	19.5 -2.6	19.0 -3.1
IN-V2 [24]	63.9	69.5 +5.5	69.9 +6.0	61.4 -2.5	61.6 -2.4

the entire extended learning process. While successful families maintain strong positive gains over the No-Transfer baseline, the failure families remain consistently negative even with substantially longer training. This confirms that the observed success/failure split is robust to extended training and that the 20-epoch results in Fig. 1 are sufficient to capture this boundary.

Dataset Robustness. We then test whether the observed failure is merely an artifact of the ImageNet-1K transfer dataset. Table 1 reports the performance of Attention Distillation on three fine-grained iNaturalist variants [20] under the standard 20-epoch protocol. The same success/failure split appears consistently across all three year splits, matching the ImageNet-1K results. This shows that the failure of certain families is robust across different datasets.

Out-of-Distribution Robustness. This observed success/failure split also persists under the out-of-distribution evaluation. Table 2 shows that on four ImageNet-1K distribution-shift benchmarks [21–24], Attention Distillation preserves the same positive/negative pattern observed in the in-distribution settings. This further confirms that the family-consistent failure of Attention Transfer is robust to substantial distribution shift.

4 Localizing the Failure

Having established the robust failure of Attention Transfer for specific teacher families, we now investigate exactly *where* the transferred pre-trained weights harm the student. We diagnose this problem systematically through the three-level analysis below.

Component-Level Decomposition. We first ask whether the family-specific split is driven by a particular ViT component or reflects a generic difference in the usefulness of the pre-trained weights. To test this, we selectively initialize the attention and MLP blocks from representative teachers under a component-level decomposition setup with full fine-tuning. As shown in Table 3, MLP-only initialization is consistently beneficial for both the success teacher (DeiT-S) and the failure teachers (DINOv2-S and CLIP-B), aligning with the findings in [25]. In contrast, Attention-only initialization preserves the success/failure split. This shows that the failure families do not suffer from a generic inability to benefit from the pre-trained weights as a whole. Instead, the failure is consistently localized to the attention pathway, motivating a finer-grained analysis of where this harm actually appears within attention itself.

Layer-wise Decomposition. We therefore ask whether the harmful effect of Attention Transfer is concentrated at a particular depth. To test this, we transfer attention from the top or bottom k layers

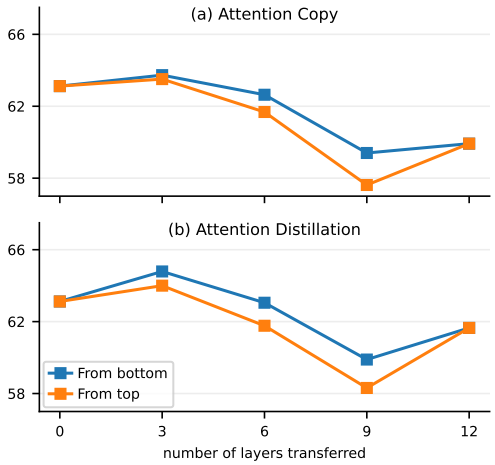

Figure 3: **Layer-wise decomposition.** We transfer attention from the top or bottom k layers of DINOv2-S under both Attention Copy and Attention Distillation. In contrast to the monotonic gains in [8], no layer subset reproduces success-family gains, and harmful transfer appears across every transfer depth for this failure teacher.

Table 3: **Component-level decomposition.** We selectively initialize attention (Attn.) and MLP blocks from three representative teachers and fine-tune end-to-end. ✓ denotes initialized components, and blank cells denote random initialization. Attention-only preserves success/failure split, while MLP-only is uniformly strong across all three families.

Attn.	MLP	DeiT-S	DINOv2-S	CLIP-B
		66.0	65.3	70.3
✓		71.7 +5.7	63.7 -1.6	69.4 -0.9
	✓	77.7 +11.7	69.2 +3.9	74.7 +4.4
✓	✓	79.3 +13.3	71.0 +5.7	80.1 +9.8

Table 4: **QKV decomposition.** We replace the default full-attention transfer with subset transfer over Q, K, and V under both methods. All alternatives improve performance on the success family but remain negative and even worse on the failure family, indicating the failure does not lie in any specific attention subcomponent.

	DeiT-S (NoT: 63.1)		DINOv2-S (NoT: 63.1)	
	Copy	Distill	Copy	Distill
Q	78.1 +15.0	77.0 +13.9	57.9 -5.2	59.0 -4.1
K	78.7 +15.6	76.7 +13.6	57.4 -5.7	58.7 -4.4
V	78.4 +15.3	78.0 +14.9	47.3 -15.9	54.9 -8.3
Attn.	75.3 +12.2	75.6 +12.4	59.9 -3.2	61.7 -1.5

of one representative failure teacher, DINOv2-S, under both Attention Copy and Attention Distillation. As shown in Fig. 3, no layer subset recovers the positive gains observed for success families. Instead, performance remains below the No-Transfer baseline across all tested transfer depths, in clear contrast to the observed monotonic-gain pattern for success teachers in [8]. This rules out a single problematic depth as the source of failure and suggests that the failure is distributed broadly across the transferred attention patterns.

QKV Decomposition. We finally ask whether the failure can be attributed to a single attention subcomponent. Across both Attention Copy and Attention Distillation, Table 4 shows that transferring any of Q, K, or V alone remains beneficial for the success teacher DeiT-S, consistent with the findings in [8]. However, the same single-component transfer is harmful for the failure teacher DINOv2-S, with V producing the largest drop. Together with the previous layer-wise analysis, this further confirms that the failure is not confined to one specific layer or one isolated attention subcomponent, but is distributed broadly throughout the overall attention pathway.

This directly motivates the next question: why does transferred attention remain functional for some teacher families, yet become ineffective for others under the same standard student architecture?

5 Architectural Compatibility Is the Key

Having localized the failure to the attention pathway, one question remains: what determines whether transferred attention is functional across teacher families? We address this below, starting with a simple native-architecture rescue.

5.1 Native-architecture Rescue

Motivation. Results in [8] suggest that transferring *only* the attention can recover the full benefit of the teacher's pre-trained weights under a standard teacher-student architecture. However, we revisit this claim in Sec. 3 to find that this does not hold universally, and systematically localize the failure to the attention pathway in Sec. 4, further confirming that this failure is broadly distributed across attention layers and Q/K/V terms. This leads us to hypothesize that the missing factor is *architectural compatibility* in the default protocol: transferred attention remains usable only when the student preserves the architectural components under which the teacher originally learned its attention patterns. Concretely, each failure family in Fig. 1 contains one or more such components that are absent from the standard student architecture: `LayerScale`[4] in DINOv2 [2] and DINOv2-wr [17], `PreLayerNorm` in CLIP [4], and `LayerScale` together with `RelativePositionBias` in BEiTv2 [14]. These native components define the scaling, normalization, and positional context under which the teacher's attention was learned. Without them, transferring attention into the standard student may instantiate the teacher's routing patterns through a mismatched computation pathway.

Experimental Protocol. To test this hypothesis, we augment the student architecture by adding only the teacher-native missing components, while keeping all other implementation details identical to Sec. 2.2. The inserted components are initialized *randomly* and trained jointly with the student,

[4] `LayerScale` was originally introduced in CaiT [26] to stabilize the training of very deep ViTs, and is not adopted by the ViT implementations of [1, 9], which is used in the standard teacher-student architecture.

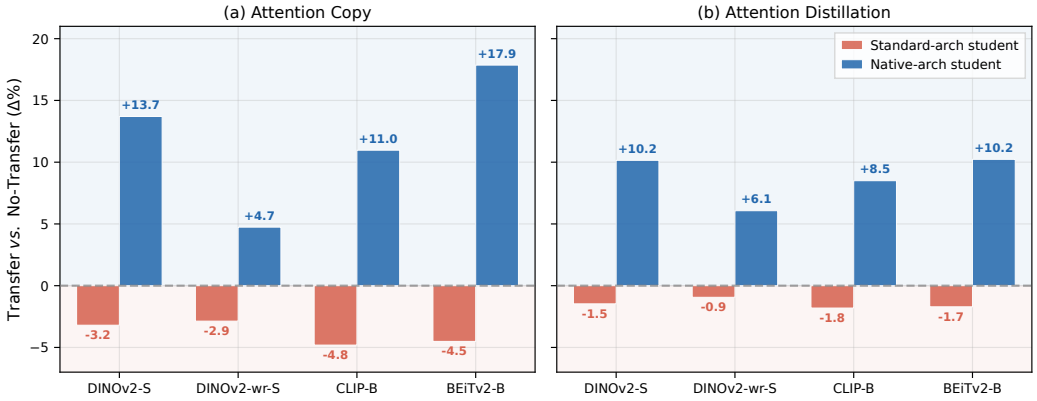

Figure 4: **Architectural compatibility rescues ineffective Attention Transfer.** For teachers from the failure families, we report Δ accuracy (%) relative to the corresponding No-Transfer baselines for both Attention Copy and Attention Distillation. The native-arch student is augmented by adding only the teacher-native components missing from the standard student in a randomly initialized state, with no pre-trained weights loaded. Across all failure families and both transfer methods, standard-arch transfer remains harmful, while native-arch transfer consistently flips to positive gains, confirming that the failure is driven by teacher-student architectural mismatch.

with no pre-trained weights loaded. We evaluate this fix under both Attention Copy and Attention Distillation for all 4 representative failure teachers. For reference, we additionally include the native-architecture No-Transfer (NoT) baseline, where the augmented student is randomly initialized and trained from scratch.

Results & Analysis. Fig. 4 demonstrates a clear and consistent performance flip for both Attention Transfer methods across the 4 failure families: transfer is harmful with the standard student architecture, while beneficial once teacher-native components are added. However, one natural concern is that these gains could come from the added components' own capacity, rather than from restored compatibility with the teacher's attention patterns. To rule this out, we compare NoT baselines for standard and native architectures. As shown in Table 5, native-arch NoT baselines yield worse performance than standard-arch ones for all failure families, indicating that the added components do not by themselves improve performance. The rescue effect therefore comes specifically from restoring architectural compatibility and unlocking effective Attention Transfer.

Table 5: **No-Transfer baseline comparison.** We compare from-scratch No-Transfer baselines between standard student and native-arch student. The native-arch baseline is lower than the standard-arch one, confirming the rescue is not driven by the added components' own capacity.

	DINOv2-S	CLIP-B	BEiTv2-B
Standard-arch	63.1	69.0	69.0
Native-arch	60.1 -3.0	66.7 -2.3	66.3 -2.7

5.2 Why Architectural Mismatch Causes Failure

To explain *why* this rescue is effective, we further test the architectural compatibility hypothesis directly at the attention level: given that the teacher's pre-trained weights computed attention under its native components, a standard student lacking these components utilizes a different attention computation pathway and would be likely to fail to match the teacher's attention. We measure this via per-layer KL divergence between the transferred student's attention and the teacher's native attention, for both standard-arch and native-arch students under Attention Distillation. As shown in Fig. 5, the standard-arch student matches the early layers almost perfectly with near-zero KL across all three failure families, but exhibits a sharp divergence spike in the last three layers. By contrast, the native-architecture student remains consistently low and smooth across all layers, without any collapse for transferring attention.

These results further explain the previously observed failure families: architectural mismatch in standard Attention Transfer leads to inconsistent late-layer attention learning, whereas adding native components resolves this compatibility issue and can then recover the transfer gains. Moreover, this also clarifies the observation in Fig. 3: From-top subsets always include the unmatched final layers, whereas From-bottom subsets stay in the matchable early-layer region, explaining why From-top consistently underperforms From-bottom. More results are given in Sec. B.1.

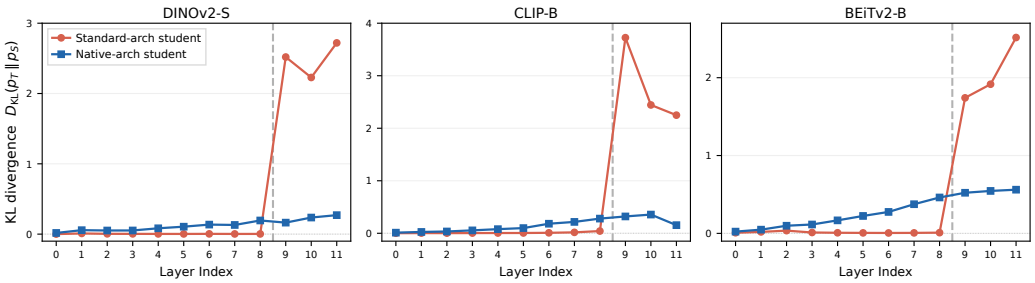

Figure 5: **Layer-wise teacher-to-student KL divergence.** We plot per-layer KL divergence between the pre-trained teacher's attention maps and the transferred student's attention maps on ImageNet-1K under Attention Distillation. A clear spike of the standard-arch student at late layers can be observed, while the native-arch student consistently stays near-zero. This shows Attention Transfer can match teacher attention only when the architecture is compatible, further explaining the rescue in Fig. 4.

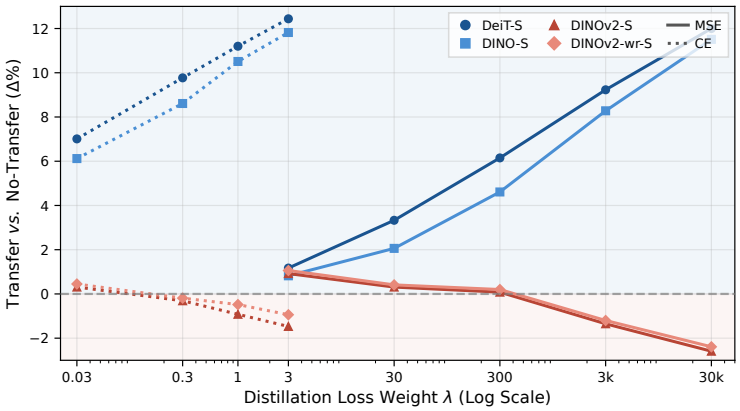

Figure 6: **Ablation on behaviors of MSE loss *vs*. default CE loss.** We compare behaviors of different transfer losses in Attention Distillation with varying λ. Same family-level trends can be observed under the λ-sweep. For success teachers, both losses yield monotonic gains with increasing λ; while for failure teachers, both losses cause larger drops with increasing λ.

Together, these results support architectural compatibility as the primary mechanism behind Attention Transfer's effectiveness: attention is sufficient to recover the full benefit of the teacher's pre-trained weights *only* when the student architecture provides the same attention routing as the teacher's.

6 Ruling Out Alternative Explanations

By establishing architectural compatibility as the primary mechanism in the previous discussion, we now ask whether simpler explanations could account for the observed failure. We therefore investigate three natural alternative explanations below.

6.1 Not the Transfer Loss

One natural explanation for the failure is the usage of a potential unsuitable loss in Attention Transfer. Specifically, Cross-Entropy (CE) loss is used by default for Attention Distillation, which might not be generally effective for transferring all families of pre-trained weights, and switching to a different transfer loss could rescue the failure ones. We therefore first use MSE loss as one alternative to study the behaviors of different losses with varying λ[5] on representative success and failure teachers.

As shown in Fig. 6, MSE exhibits the same family-level trends as CE when the two losses' magnitudes are matched by adjusting λ. For success teachers (DeiT-S, DINO-S), both losses produce monotonic improvements as λ increases, and MSE approaches the substantial same level of performance as CE when sufficient magnitude is given by large λ. For failure teachers (DINOv2-S, DINOv2-wr-S), both losses instead produce monotonic degradation as λ increases. This shows that increasing the

[5]We follow [8] to denote λ as the weight of the distillation loss in the overall training objective of Attention Distillation.

Table 6: **Counter-examples for potential pre-training recipe properties.** We list the candidate properties below from three main categories. For each candidate, we identify either one success teacher sharing the same property as the failure teacher, or one teacher without this property but still fails. None of the tested single-factor recipe properties predicts the family-level failure boundary.

Hypothesis	**Failure**	**Counter-example**
Pre-training signal	DINOv2 (Self-Distillation)	DINO (Self-Distillation)
	CLIP (Contrastive)	SigLIP2 (Contrastive)
	BEiTv2 (Masked Image Modeling)	iBOT/MAE (Masked Image Modeling)
Data source	DINOv2 (Non-ImageNet Data)	SAM (Non-ImageNet Data)
	CLIP (Multi-Modal Data)	SigLIP2 (Multi-Modal Data)
Special attribute	DINOv2 (with Attention Sink)	DeiT (with Attention Sink)
	DINOv2 (patch $14 \rightarrow 16$ to fit standard-arch)	CLIP (native patch 16, also fails)

Attention Transfer signal helps success families but amplifies harm for failure families, regardless of whether the transfer loss is CE or MSE. We further verify this pattern across additional losses in Sec. B.6, all demonstrating the same pattern as MSE. These results rule out the inadequate choice of transfer loss as the primary explanation for the observed failure.

6.2 Not the Pre-training Recipe

Another natural explanation is that this family-dependent failure may stem from common properties of the teachers' pre-training recipes. We test several candidate properties in the pre-training below. Specifically, we summarize three main categories of hypotheses, including different pre-training signals, different data sources, and other known special attributes such as attention sinks and patch interpolation. As shown in Table 6, for each candidate explanation from the failure families, we can identify at least one counter-example with that property from the success families, or one teacher that lacks it but still fails. Overall, none of these single-factor recipe properties predicts the observed family-level failure boundary. This rules out pre-training recipe differences as the primary explanation. Detailed discussion of these recipe properties is provided in Sec. B.7, together with further analysis on controlled experiments that would fully isolate the pre-training recipe.

6.3 Not the Attention Map Quality

The third natural explanation is that the failure families simply have attention maps of poorer intrinsic quality, thus a student could hardly benefit from them. Prior evidence already argues against this: the attention of MAE is known to exhibit strong locality with weak semantic structure [27], yet MAE transfers successfully, while strong semantic localization is a headline property of DINOv2 [2], yet DINOv2 fails most severely. To test this directly, we measure the quality of attention maps with the object localization protocol of SimPool [28]: MaxBoxAccV2 [29] on 2,000 ImageNet-1K validation images with ground-truth boxes. More details are given in Sec. B.4.

We first evaluate the pre-trained teachers before the transfer. As shown in Table 7, the success teacher DINO-B scores higher than both failure teachers, while the other success teacher DeiT-B scores below them: the quality ranking is interleaved with the success/failure split, and the same holds at every divergence-spike layer as shown in Sec. B.4. The intrinsic quality of the teacher's attention maps therefore does not predict the family-dependent failure.

Furthermore, the proposed native-arch rescue does not work through attention map quality. Under Attention Copy, the student's attention maps are the teacher's by construction, so we compare the students under Attention Distillation, where the student learns its own attention. As shown in Table 8, the rescue has no consistent effect direction on the attention map quality:

Table 7: **Attention map quality comparison before Attention Transfer.** We report MaxBoxAccV2 (%) of each teacher's final-layer attention under the SimPool [28] localization protocol. The quality ranking is interleaved with the success/failure split.

	DINO-B	DINOv2-B	CLIP-B	DeiT-B
MaxBoxAccV2	79.9	69.8	68.9	51.6

Table 8: **Attention map quality comparison of transferred students w/ and w/o the rescue.** We report MaxBoxAccV2 (%) of the transferred students under Attention Distillation for three failure teachers. The rescue shows no consistent effect direction on the attention map quality.

	DINOv2-S	CLIP-B	BEiTv2-B
w/o rescue	76.1	72.0	68.2
w/ rescue	83.7 +7.6	67.9 -4.1	51.6 -16.6

it raises the measured quality under DINOv2-S but lowers it for CLIP-B and BEiTv2-B. In contrast, this rescue can reverse the transfer failure for all three weights as shown in Fig. 4. This cleanly separates the two explanations: the rescue works through restoring architectural compatibility, rather than through changing the attention map quality. These results thus rule out attention map quality as the primary explanation for the observed failure.

After ruling out the inadequate choice of transfer loss, differences in pre-training recipe, and the quality of attention maps, we find that none of these alternatives accounts for the family-level failure boundary. Taken together with Sec. 5, these analyses further support architectural mismatch as the primary mechanism behind the failure, further refining the prevailing understanding of attention in ViT representations: attention is sufficient *only* when the student architecture matches the teacher.

7 Related Work

ViT Variants. Since the Transformer was introduced to vision by [1], a diverse landscape of pre-training paradigms has been proposed, including supervised training [9, 26, 30, 31], self-distillation [10, 2, 13, 3], contrastive learning [32, 33, 11], multi-modal pre-training [4, 34, 35, 15], masked image modeling [12, 36, 14, 37], and large-scale segmentation pre-training [16, 38, 39], with further analyses trying to understand the underlying mechanisms behind their effectiveness [5–7, 40]. Several modern ViT families introduce architectural components beyond the standard design, including LayerScale in [26, 2], PreLayerNorm in [4], learnable relative position bias in [36, 14], and RoPE in [41, 3]. A separate line of work has analyzed the structural role of attention itself in ViTs, aiming to mitigate attention sinks via register tokens [17, 42] or sliding windows [43]. However, the role of attention design in determining ViT effectiveness remains fundamentally unclear.

Attention Transfer. The term *Attention Transfer* was originally coined for distilling spatial activation maps in CNNs [44], while related work in language modeling demonstrated that self-attention distributions alone can serve as a powerful distillation signal [45]. This idea was later extended to the vision domain: to guide self-supervised ViT students [46], or to initialize ViT attention layers [47]. Building on this line, [8] further demonstrated that transferring only the model's attention patterns is sufficient to recover the full benefit of its pre-trained weights. Some contemporary directions have explored richer feature transfer beyond attention [48–53], initializing smaller ViTs from larger ones [54], or transferring across attention operators [55–57]. While these methods motivate Attention Transfer's appeal as a minimal-signal alternative, evaluations in [8] are limited to a few families compared to the growing diverse landscape of ViTs.

8 Conclusion

In this work, we revisit the recent claim that Attention Transfer [8] is sufficient to recover the full benefit of the pre-trained weights for Vision Transformers. On a comprehensive model zoo of 20 teachers from 11 well-known ViT families, we reveal a hidden validity boundary: 4 families consistently fail under the standard teacher-student architecture, falling up to 5.1% below the from-scratch no-transfer baseline, with the failure persisting across model sizes, extended training durations, different transfer datasets, and out-of-distribution evaluations. Through a three-level localization analysis and a simple native-architecture rescue, we identify architectural mismatch between teacher and student as the primary mechanism behind this failure, and further rule out the inadequate choice of transfer loss, differences in pre-training recipe, and the quality of attention maps as alternative explanations. Our findings refine the prevailing understanding of attention in ViT representations: attention is not architecture-agnostically transferable, and is sufficient *only* when the student architecture preserves the same attention routing as the teacher's. We hope our explorations, together with the corresponding analyses, could encourage the community to revisit the underlying mechanisms of Attention Transfer and further unlock the potential of this transfer paradigm.

Acknowledgments and Disclosure of Funding

This research is supported by the NSF of China (Grant No. 62336003, 12371510) and the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G12), and is conducted in collaboration with ST Engineering Geo-Insights Pte Ltd. The authors sincerely thank the anonymous Reviewers and Area Chairs for their valuable feedback and constructive comments, and declare there are no competing interests related to this research.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 1, 3, 6, 10

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2, 4, 6, 9, 10, 15, 16

[3] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *TMLR*, 2026. 10

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 4, 6, 10, 15, 16

[5] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *CVPR*, pages 7486–7496, 2023. 1, 10

[6] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *ICLR*, 2023.

[7] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *ICCV*, pages 1934–1943, 2023. 1, 10

[8] Alexander C Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. On the surprising effectiveness of attention transfer for vision transformers. In *NeurIPS*, pages 113963–113990, 2024. 2, 3, 4, 5, 6, 8, 10, 15, 18, 20, 21

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 2, 3, 4, 6, 10, 15, 16

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 4, 10, 15, 16

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 4, 10, 15, 16

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 4, 10, 15, 16

[13] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 4, 10, 16

[14] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2, 4, 6, 10, 15, 16

[15] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 4, 10, 15, 16

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 4, 10, 15, 16

[17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2, 4, 6, 10, 15, 16, 19

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *CVPR*, pages 248–255, 2009. 4, 16

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4, 16

[20] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 5

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 5

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 5

[23] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 5

[24] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 5

[25] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *EMNLP*, pages 5484–5495, 2021. 5

[26] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, pages 32–42, 2021. 6, 10

[27] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *CVPR*, 2023. 9

[28] Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it SimPool: Who said supervised transformers suffer from attention deficit? In *ICCV*, 2023. 9, 17

[29] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020. 9, 17

[30] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *ECCV*, pages 497–515, 2022. 10

[31] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, pages 516–533, 2022. 10

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 10

[33] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 10

[34] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 10

[35] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 10

[36] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 10

[37] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 10

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. 10

[39] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. 10

[40] Huaiyuan Qin, Muli Yang, Siyuan Hu, Peng Hu, Yu Zhang, Chen Gong, and Hongyuan Zhu. Beyond instance consistency: Investigating view diversity in self-supervised learning. *TMLR*, 2025. 10

[41] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, pages 289–305, 2024. 10

[42] Nick Jiang, Amil Dravid, Alexei A Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *NeurIPS*, 2025. 10

[43] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024. 10, 19

[44] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 10

[45] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, pages 5776–5788, 2020. 10

[46] Kai Wang, Fei Yang, and Joost van de Weijer. Attention distillation: Self-supervised vision transformer students need more guidance. In *BMVC*, 2022. 10

[47] Asher Trockman and J. Zico Kolter. Mimetic initialization of self-attention layers. In *ICML*, pages 34456–34468, 2023. 10

[48] Jiawei Fan, Chao Li, Xiaolong Liu, and Anbang Yao. ScaleKD: Strong vision transformers could be excellent teachers. In *NeurIPS*, pages 63290–63315, 2024. 10

[49] Asher Trockman and J. Zico Kolter. Mimetic initialization of mlps. *arXiv preprint arXiv:2602.07156*, 2026.

[50] Penghao Wang, Yuhao Zhou, Mengxuan Wu, Panpan Zhang, Zhangyang Wang, and Kai Wang. Data efficient any transformer-to-mamba distillation via attention bridge. *arXiv preprint arXiv:2510.19266*, 2025.

[51] Huiyuan Tian, Bonan Xu, and Shijian Li. From per-image low-rank to encoding mismatch: Rethinking feature distillation in vision transformers. In *ICML*, 2026.

[52] Ying Peng, Hongsen Ye, Changxin Huang, Xiping Hu, Jian Chen, and Runhao Zeng. Revisiting cross-architecture distillation: Adaptive dual-teacher transfer for lightweight video models. In *AAAI*, pages 8367–8375, 2026.

[53] Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Perspective-aware teaching: Adapting knowledge for heterogeneous distillation. In *ICCV*, pages 4178–4187, 2025. 10

[54] Zhiqiu Xu, Yanjie Chen, Kirill Vishniakov, Yida Yin, Zhiqiang Shen, Trevor Darrell, Lingjie Liu, and Zhuang Liu. Initializing models with larger ones. In *ICLR*, 2024. 10

[55] Huaiyuan Qin, Muli Yang, Gabriel James Goenawan, Shiqi Huang, Min Kass Chong, Wahyu Wiratama, Peng Hu, Chen Gong, Wu Liu, Xi Peng, et al. Copy the same, distill the difference: Initializing linear vision transformers. *arXiv preprint arXiv:2609.35745*, 2026. 10

[56] Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *ICLR*, pages 53551–53580, 2024.

[57] Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Frederick Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. Lolcats: On low-rank linearizing of large language models. In *ICLR*, pages 46200–46253, 2025. 10

[58] Shashanka Venkataramanan, Amir Ghodrati, Yuki M Asano, Fatih Porikli, and Amirhossein Habibian. Skip-attention: Improving vision transformers by paying less attention. In *ICLR*, 2024. 17

Attention Transfer Is Not Universally Effective for Vision Transformers

Technical Appendices and Supplementary Material

Contents

1 Introduction **1**

2 Preliminaries **3**
- 2.1 Attention Transfer . . . 3
- 2.2 Experimental Setup . . . 4

3 Attention Transfer Is Not Universally Effective **4**
- 3.1 Main Results . . . 4
- 3.2 Robustness of the Failure . . . 4

4 Localizing the Failure **5**

5 Architectural Compatibility Is the Key **6**
- 5.1 Native-architecture Rescue . . . 6
- 5.2 Why Architectural Mismatch Causes Failure . . . 7

6 Ruling Out Alternative Explanations **8**
- 6.1 Not the Transfer Loss . . . 8
- 6.2 Not the Pre-training Recipe . . . 9
- 6.3 Not the Attention Map Quality . . . 9

7 Related Work **10**

8 Conclusion **10**

A Additional Implementation Details **15**
- A.1 Details of Model Zoo . . . 15
- A.2 Details of Experimental Setup . . . 15

B Additional Results and Analysis **15**
- B.1 Additional Analysis on Architectural Mismatch . . . 15
- B.2 Additional Analysis on Native-architecture Rescue . . . 16
- B.3 Additional Analysis on Late-layer Interventions . . . 16
- B.4 Additional Analysis on Attention Map Quality . . . 17
- B.5 Additional Analysis on Relation to Knowledge Distillation . . . 18
- B.6 Additional Analysis on Transfer Loss . . . 18
- B.7 Additional Analysis on Pre-training Recipe . . . 18
- B.8 Additional Analysis on Sensitivity of Attention Distillation Failure . . . 20
- B.9 Additional Analysis on Statistical Significance . . . 20

C Limitations & Scope **20**

D Broader Impacts **21**

A Additional Implementation Details

We provide further implementation details to support reproducibility of our experimental setup below.

A.1 Details of Model Zoo

Table B summarizes all 20 pre-trained weights from 11 ViT families used in our evaluations, covering three model sizes (Tiny, Small, Base) where official checkpoints are available. The benchmark spans the major pre-training paradigms in recent ViT representation learning, including supervised training [9], self-distillation [10, 2, 17], contrastive learning [11], multi-modal pre-training [4, 15], masked image modeling [12–14], and large-scale segmentation pre-training [16].

A.2 Details of Experimental Setup

Table C summarizes the training recipes used for Attention Copy and Attention Distillation in our evaluations. We follow the protocol in [8] to ensure direct comparability with their findings, with the following deliberate modifications for our broader evaluations: (i) we use 20-epoch training for the main benchmark and further confirm in Fig. 2 that this is sufficient to reveal the family-level success/failure boundary; (ii) all reported results are averaged over 3 random seeds to ensure statistical significance. All hyper-parameters are held identical across all 20 teachers, ensuring no per-family tuning artifacts. All experiments are conducted on NVIDIA A100 GPUs.

B Additional Results and Analysis

B.1 Additional Analysis on Architectural Mismatch

Layer-wise Divergence. The teacher-to-student divergence in Fig. 5 is computed with per-layer attention maps from both the transferred student and the pre-trained teacher. Unlike [8], which compares different transfer methods against fine-tuned models using Jensen-Shannon (JS) divergence, our focus is to assess whether the transferred student can match the pre-trained teacher's attention for both standard-arch and native-arch students. We use Kullback-Leibler (KL) divergence for this assessment under Attention Distillation and additionally report JS divergence in Fig. A. The JS curves demonstrate the same pattern as KL curves, further supporting that the architectural mismatch in standard Attention Transfer leads to inconsistent late-layer attention learning.

CLS *vs*. Patch Decomposition. The divergences discussed above are computed by averaging over all 197 query tokens, which include one CLS token and 196 patch tokens. Since the late layers are also where the CLS token becomes most important for classification, we further decompose the per-layer divergence into the CLS query row and the patch query rows for further investigation. We test this with the representative failure teacher, DINOv2-S, with the standard-arch student under Attention Distillation.

Table A: **CLS *vs*. patch decomposition of the late-layer divergence.** We report the per-layer KL and JS divergence between the teacher and standard-arch student under Attention Distillation on DINOv2-S.

Layer	0-8 (avg)	9	10	11
KL on CLS	< 0.01	1.84	2.40	1.27
KL on Patch	< 0.01	2.52	2.23	2.73
JS on CLS	< 0.001	0.34	0.40	0.25
JS on Patch	< 0.001	0.33	0.35	0.40

As shown in Table A, both the CLS row and the patch rows exhibit near-zero divergence in the early layers and high divergence in the last three layers, with comparable magnitudes, and the patch rows diverging more strongly in two of the three spike layers. The late-layer mismatch therefore exists in both the global routing and the local routing, rather than being concentrated in one type of token, which is consistent with our hypothesis: the missing native component modulates the routing of all tokens, so the induced mismatch is expected to be distributed across the entire attention computation.

Moreover, based on these decomposition results, we expect that replacing the CLS readout with global average pooling in the standard student would not fully resolve the failure, since the patch-level incompatibility inside the attention computation remains unchanged. Given the training cost of this diagnostic, we would like to leave the empirical validation of this expectation as future work.

Table B: **Pre-trained Model Zoo.** We present 20 pre-trained weights from 11 well-known ViT families used in our Attention Transfer evaluations. All weights are with the patch size of 16, while for families only with patch size of 14 [2, 17], we interpolate their patch embeddings from 14 to 16.

Family	Size	Pre-training Data	Pre-training Paradigm	Native-Arch Components
Success families				
DeiT [9]	T, S, B	ImageNet-1K [18]	Supervised	–
DINO [10]	S, B	ImageNet-1K [18]	Self-Distillation	–
MoCov3 [11]	T, S, B	ImageNet-1K [18]	Contrastive (momentum)	–
MAE [12]	T, B	ImageNet-1K [18]	Masked Image Modeling	–
iBOT [13]	S, B	ImageNet-1K [18]	Self-Distillation (masked)	–
SigLIP2 [15]	B	WebLI [15]	Multi-Modal Contrastive	–
SAM [16]	B	SA-1B [16]	Segmentation Pre-training	–
Failure families				
DINOv2 [2]	S, B	LVD-142M [2]	Self-Distillation + MIM	`LayerScale`
DINOv2-wr [17]	S, B	LVD-142M [2]	+ Register Tokens	`LayerScale`
CLIP [4]	B	WIT-400M [4]	Multi-Modal Contrastive	`PreLayerNorm`
BEiTv2 [14]	B	ImageNet-1K [18]	Masked Image Modeling	`LayerScale`, `RelativePositionBias`

Table C: **Training recipes for Attention Copy and Attention Distillation.**

Config	Attention Copy	Attention Distillation
Optimizer	AdamW [19]	AdamW [19]
Base Learning Rate	1e-3	1e-4
Weight Decay	0.05	0.3
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.95$
Layer-wise LR Decay	0.75	–
Batch Size	1024	1024
LR Schedule	cosine decay	cosine decay
Warm-up Epochs	5	5
Epochs	20	20
Augmentation	RandAug (9, 0.5)	RandAug (9, 0.5)
Label Smoothing	0.1	0.1
Mixup	0.8	0.8
Cutmix	1.0	1.0
Drop Path	0	0.1
EMA	0.9999	0.9999
Distillation Loss	–	Cross-Entropy (default)
Distillation Loss Weight λ	–	3
Random Seeds	3	3

B.2 Additional Analysis on Native-architecture Rescue

Table D: **Per-component decomposition of the native-arch rescue on BEiTv2-B.**

Δ Acc	Standard	+`LayerScale`	+`RPB`	+Both
Copy	-4.5	-2.1	-1.3	+17.9
Distill	-1.7	-1.3	-0.8	+10.2

The native-architecture rescue in Fig. 4 provides results for adding all the teacher-native components missing from the standard student. Specifically, three of the four failure families have only one single native component: the native-arch student adds only `LayerScale` for DINOv2 and DINOv2-wr, and only `PreLayerNorm` for CLIP. The only multi-component case is BEiTv2, whose pre-trained weights natively contain both `LayerScale` and `RelativePositionBias`. We therefore further decompose the rescue experiment for BEiTv2-B by evaluating two additional students that add only one of the two native components. As shown in Table D, each single component alone only reduces the harm but never reverses the failure, whereas adding both flips the transfer to strongly positive gains under both transfer methods. This is consistent with our hypothesis that transferred attention remains effective only when the student preserves the full architectural context under which the teacher's attention was learned.

B.3 Additional Analysis on Late-layer Interventions

Motivated by the late-layer divergence spike in Fig. 5 and Fig. A, we ask whether the failure can be fixed by changing the distillation scheme itself, rather than restoring native components as in the native-arch rescue. We evaluate three interventions on the failure teacher DINOv2-S under the

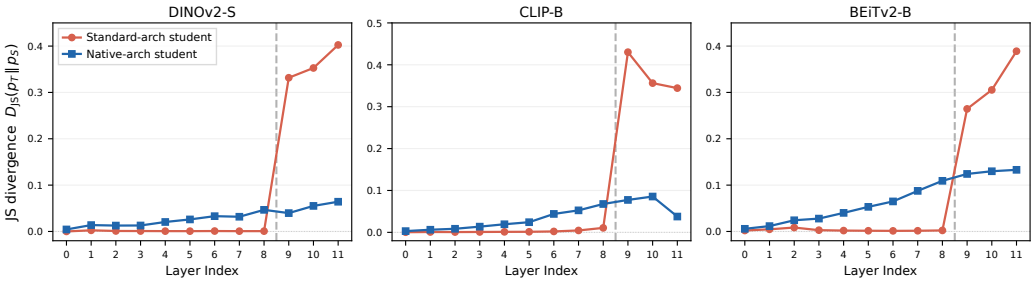

Figure A: **Layer-wise teacher-to-student JS divergence.** We plot per-layer JS divergence between the pre-trained teacher's attention maps and the transferred student's attention maps on ImageNet-1K under Attention Distillation. The qualitative pattern is identical to that in Fig. 5: standard-arch students show a sharp late-layer spike across all three failure families (DINOv2-S, CLIP-B, BEiTv2-B), while native-arch students stay low and smooth.

standard student: *From-bottom* transfers only the first 9 layers, leaving the last 3 layers free to learn their own attention, as in Fig. 3. *Reuse*, inspired by Skip-Attention [58], transfers the first 9 layers as usual while the last 3 layers reuse the teacher's layer-8 attention map, which tests whether the harmful part of the transfer is the late-layer attention content itself. *Reweight* retains supervision on all 12 layers but calibrates the distillation weight of the last 3 layers by w, aiming to test whether the problematic layers are merely under-supervised.

As shown in Table E, none of these interventions recovers the transfer benefit. Under Attention Copy, reusing early attention removes most of the harm but recovers none of the benefit, leaving the result at the No-Transfer level, whereas the native-arch rescue achieves a substantial improvement with the same teacher maps. Under Attention Distillation, reusing early attention becomes even more harmful than the default full transfer, as it forces the student to learn inappropriate token routing at the wrong depth, and strengthening the supervision exactly on the problematic layers makes the failure worse still. Together, these results indicate that avoiding the late-layer targets cannot effectively recover the transfer benefit, while only restoring the native architectural components does. This further supports our interpretation that the architectural mismatch is the key to the observed failure, rather than the late-layer attention content alone.

Table E: **Late-layer interventions within the standard-arch on DINOv2-S.** We report Δ accuracy (%) relative to the corresponding No-Transfer baselines. None of the late-layer interventions recovers the transfer benefit, while only the native-arch rescue does.

Δ	Copy	Distill
Full transfer (default)	-3.2	-1.5
From-bottom, 9 layers	-3.7	-3.2
Reuse	+0.8	-5.4
Reweight, $w = 10$	–	-8.6
Native-arch rescue	+13.7	+10.2

B.4 Additional Analysis on Attention Map Quality

We provide more details on the attention map quality measurement used in Sec. 6.3 and further extend it to per-layer measurement. Following the object localization protocol of SimPool [28], we compute MaxBoxAccV2 [29] on the first 2,000 ImageNet-1K validation images with ground-truth boxes. The CLS-to-patch attention map of a given layer is averaged over heads, then bilinearly upsampled to the input resolution with min-max normalization. For each threshold $\tau \in \{0, 0.01, \ldots, 0.99\}$, the map is binarized and every connected component is converted into a box. An image is counted as correctly localized at IoU threshold δ if any estimated box overlaps any ground-truth box with IoU $\geq \delta$, MaxBoxAccV2 is then computed by averaging the best accuracy over τ across $\delta \in \{0.3, 0.5, 0.7\}$. Teachers are evaluated with their native architectures, and students are taken from the same Attention Distillation runs as in Fig. 4.

Table F: **Per-layer attention map quality comparison.** We report per-layer MaxBoxAccV2 (%) for the last 3 divergence-spike layers. The success/failure ranking remains interleaved: this quality does not split the boundary.

	DINO-B	DINOv2-B	CLIP-B	DeiT-B
Layer 9	**84.8**	70.1	58.2	64.1
Layer 10	81.1	**82.7**	65.1	56.8
Layer 11	**79.9**	69.8	68.9	51.6

Table F extends the final-layer teacher measurement of Table 7 to each of the 3 divergence-spike layers of Fig. 5 and Fig. A. As shown, at every layer, the success/failure ordering remains interleaved and

no layer exhibits a quality split aligned with the transfer boundary. The earlier spike layers also show no consistent quality degradation separating the failure families from the successes: DINOv2-B's layer 10 is its best-scoring layer, and CLIP-B's quality gradually rises toward the final layer.

B.5 Additional Analysis on Relation to Knowledge Distillation

Our work is a mechanistic re-evaluation of the claim in [8], instead of a new distillation method: [8] claims that transferring only the attention patterns can recover the full benefit of the pre-trained weights, and our contribution is to establish *when* this claim holds and *why* it fails. Standard knowledge distillation methods are therefore *not* baselines that our conclusion competes against. The two lines of work also operate in different paradigms: standard knowledge distillation typically serves model compression, where a task-specific teacher transfers its logits or features to a smaller student, whereas Attention Transfer operates in the pre-training paradigm, transferring only the attention from a pre-trained teacher into a matched-size, randomly initialized student in order to isolate which part of the pre-trained knowledge suffices. A direct comparison is further complicated in our benchmark by the fact that most teachers are self-supervised or vision-language pre-trained models without a classifier, so logit distillation is not directly defined unless each teacher is first fine-tuned, which would mainly alter the object of study. We also note that the comparison between Attention Transfer and feature distillation is provided in [8]: distilling the features of the pre-trained teacher recovers only a small fraction of the benefit recovered by Attention Distillation within the matched-size scope.

Table G: **Attention Distillation *vs*. feature distillation comparison.** We report Δ accuracy (%) relative to the No-Transfer baselines of the corresponding students, following the feature distillation implementation in [8].

Δ	MAE-B	DINOv2-B	DINOv2-B (w/ rescue)
Attn. Distill	+8.1	-2.3	+9.5
Feat. Distill	+2.2	+1.7	+5.1

Moreover, we extend this comparison under our protocol on one success teacher and one failure teacher, together with the native-arch rescue. As shown in Table G, for the success family, feature distillation recovers only a small fraction of the Attention Transfer gain, reproducing the pattern in [8]. For the failure family, Attention Distillation harms the standard student while feature distillation still transfers positively; however, the native-architecture rescue flips Attention Distillation to outperform feature distillation again. In other words, whether Attention Transfer outperforms standard distillation is itself governed by architectural compatibility, which is precisely the validity boundary our paper establishes.

B.6 Additional Analysis on Transfer Loss

In addition to the CE loss *vs*. MSE loss discussed in Sec. 6.1, we further verify whether the observed family-level failure persists under other well-known distillation losses with distinct geometric properties. We extend the λ-sweep on the same 4 representative teachers using two additional losses: the Jensen-Shannon divergence, and the L1 distance. Plots in Fig. B together show that the family-level success/failure split observed in Fig. 6 remains consistent under both JSD and L1. For success teachers (DeiT-S, DINO-S), all four losses produce monotonic gains with increasing λ; for failure teachers (DINOv2-S, DINOv2-wr-S), all produce monotonic degradation. Notably, both the success-side gain and the failure-side harm scale monotonically with the loss's effective gradient magnitude, but the sign of the family-level split is invariant. This further confirms that the family-level failure is not an artifact of the inadequate choice of transfer loss, complementing the discussion in Sec. 6.1.

B.7 Additional Analysis on Pre-training Recipe

We provide more detailed discussion on the candidate pre-training-recipe hypotheses summarized in Table 6. For each candidate, we identify either (i) one success teacher that shares the same property with one failure teacher, or (ii) one teacher without the property that still fails.

Pre-training Signal. One natural hypothesis is that the failure families share a particular pre-training objective that the standard student cannot effectively transfer. We find that the 4 failure families cover the following three categories:

- Self-Distillation: DINOv2 and DINOv2-wr both fail under Attention Transfer, while DINO, which also uses a self-distillation paradigm based on an EMA teacher, succeeds.

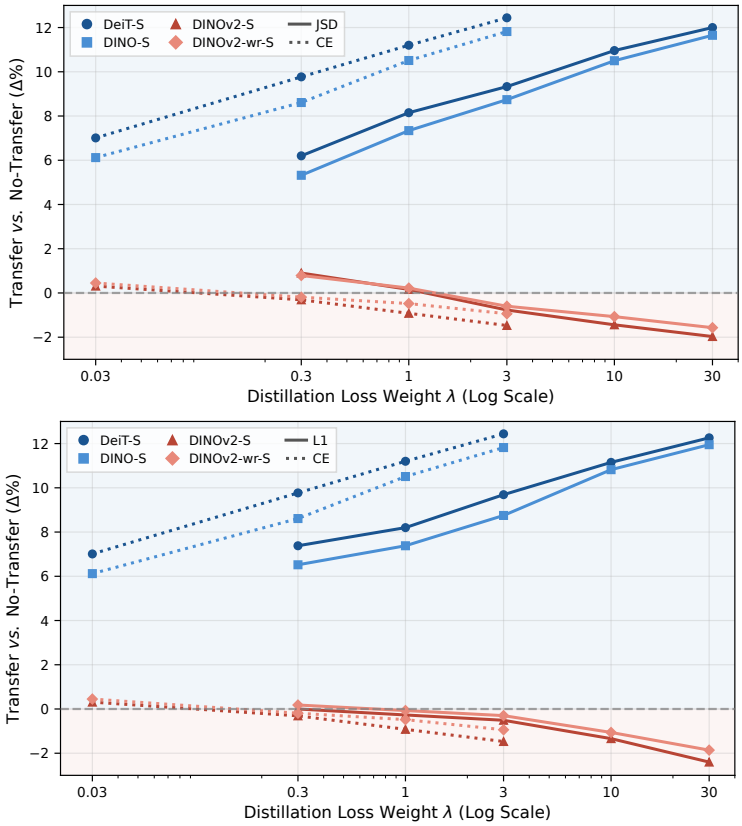

Figure B: **Ablation on behaviors of JSD loss (*top*) / L1 loss (*bottom*) *vs*. default CE loss.** We compare behaviors of different transfer losses in Attention Distillation with varying λ. Same family-level trends can be observed under the λ-sweep. For success teachers, both losses yield monotonic gains with increasing λ; while for failure teachers, both losses cause larger drops with increasing λ.

- Contrastive Multi-Modal Pre-training: CLIP fails under Attention Transfer despite being pre-trained with multi-modal contrastive learning, while SigLIP2, which also uses the same pre-training paradigm, succeeds.
- Masked Image Modeling: BEiTv2 fails under masked image modeling, while iBOT and MAE, which both perform masked-image modeling during pre-training, succeed.

The pre-training signal therefore does not predict the family-level failure boundary.

Data Source. The second hypothesis is that the failure families share a particular group of training data, because most ViT weights are trained on ImageNet-1K, with the analysis below:

- Non-ImageNet Data: DINOv2 from failure families is pre-trained on LVD-142M, a curated non-ImageNet dataset of $\sim$142M images. However, SAM is also pre-trained on one non-ImageNet dataset, SA-1B, yet succeeds.
- Multi-Modal Data: CLIP is pre-trained on a multi-modal image-text dataset, WIT-400M, and fails, while SigLIP2 is also pre-trained on a multi-modal dataset, WebLI, yet succeeds.

The pre-training data source therefore does not predict the family-level failure boundary.

Special Attribute. The third hypothesis is that the failure families share special structural properties in their attention or specific implementation tricks.

- Attention Sink: DINOv2 has been found to develop strong attention sinks during pre-training [17], which has been speculated to disrupt downstream transferability of attention. However, attention sinks are also well-documented in standard supervised ViTs such as DeiT [17, 43], yet DeiT succeeds.

- Patch-Size Interpolation: DINOv2 is officially released with patch size 14, which we adapt to patch size 16 via patch-embedding interpolation to fit the standard teacher-student architecture protocol. One may suspect that this interpolation introduces representational artifacts that cause the observed failure. However, CLIP is natively released with patch size 16 and is used with no interpolation in our setup, yet still fails under Attention Transfer.

Taken together, none of these candidate single-factor properties of pre-training recipes individually predicts the previously observed family-level failure boundary. This further supports our central claim in Sec. 5 that the family-level boundary is driven by architectural mismatch between the pre-trained teacher and the standard student, rather than by any specific aspect of the pre-training recipe.

Towards Fully Controlled Experiments. The above analysis is based on counter-examples rather than fully controlled pre-training, where such an experiment would isolate the architectural factor from the pre-training recipe at the teacher side, complementing our native-arch rescue, which isolates it at the student side. For a failure family, the controlled experiment would fix the pre-training data, objective, and all hyper-parameters, and pre-train two teachers that differ *only* in the native architectural components: one with the original native architecture, and one with the standard ViT architecture. Our hypothesis is that Attention Transfer from the standard-arch teacher into the standard student would then behave like the success families. For a success family, the controlled experiment would add the native components to the teacher under the same pre-training recipe, with the hypothesis that Attention Transfer from this modified teacher into the standard student would then degrade toward the failure pattern. Since the original pre-training data of failure families such as DINOv2 and CLIP is not publicly available, such experiments would have to be conducted on open-source data at a reduced scale. We believe this is a valuable direction and leave it for future work.

B.8 Additional Analysis on Sensitivity of Attention Distillation Failure

We further clarify the robustness of the observed failure from the following aspects: First, to ensure a fair comparison across all 20 teachers, we strictly keep all training hyper-parameters identical by following the default settings of [8], which were originally tuned to make Attention Transfer succeed; no per-family tuning is involved. Second, the family-level failure is equally reflected by Attention Copy, which involves no distillation-specific optimization at all. The same failure pattern under both Attention Transfer methods indicates that the failure is not created by distillation-specific optimization choices. Third, the transfer objective in Attention Distillation is jointly optimized with the downstream task training. Changing the learning rate, schedule, or warm-up also changes the entire optimization, including the No-Transfer baseline itself, which could also affect the effectiveness of general task optimization. For a controlled sensitivity analysis, the adjustable transfer-specific factors are therefore only the distillation loss weight and the choice of loss, both of which are swept extensively in Sec. 6.1 and Sec. B.6 to show that the family-level failure never changes. Finally, one remaining accessible robustness test is longer training, which we have examined in Fig. 2: the failure remains consistent throughout and tends to converge by 100 epochs, confirming that the negative transfer is not an artifact of insufficient optimization.

B.9 Additional Analysis on Statistical Significance

We further provide the statistical significance analysis of our main results in Fig. 1. We report the per-teacher standard deviation of Δ accuracy under both Attention Copy and Attention Distillation for all 6 pre-trained weights with model size Small across 3 random seeds in Table H.

C Limitations & Scope

Our findings are scoped to Attention Transfer protocol studied in [8]: transferring softmax attention patterns from a pre-trained ViT teacher to a randomly initialized ViT student through either Attention Copy or Attention Distillation. Within this setting, our results show that the effectiveness of Attention Transfer is governed by architectural compatibility between teacher and student. This claim concerns only softmax-to-softmax attention-pattern transfer, rather than all possible ways of reusing pre-trained ViT weights or converting attention mechanisms. Other transfer paradigms, such as full-weight initialization, feature/knowledge distillation, parameter-efficient fine-tuning, or transferring softmax attention to linear or sparse attention variants, fall outside the scope of our conclusions.

Table H: **Statistical significance analysis on main results.** We report the per-teacher standard deviation of our main results on all 6 pre-trained weights with model size Small over 3 random seeds. The standard deviations are uniformly within 0.2%, far below the magnitude of every reported Δ, confirming that the observed family-level failure boundary in Fig. 1 is robust to seed variation.

	DeiT-S	DINO-S	MoCov3-S	iBOT-S	DINOv2-S	DINOv2-wr-S
Δ Acc (Copy)	+12.2	+11.1	+4.8	+11.9	-3.2	-2.9
$\pm$ std	0.13	0.17	0.10	0.15	0.07	0.11
Δ Acc (Distill)	+12.4	+12.0	+7.5	+12.1	-1.5	-0.9
$\pm$ std	0.14	0.16	0.11	0.13	0.09	0.10

The architectural compatibility principle concerns whether the student ViT preserves the architectural context under which the teacher learned its attention patterns. For the failure families identified in our benchmark, this context is captured by concrete teacher-native components, including `LayerScale`, `PreLayerNorm`, and `RelativePositionBias`, which are introduced by different ViT variants. Our native-architecture rescue shows that restoring these components in randomly initialized form is sufficient to make the transferred attention usable, so we interpret them as architectural scaffolding rather than additional pre-trained knowledge. Our experiments use matched-size teacher–student pairs, and the findings should therefore be read as characterizing architectural compatibility under matched-capacity transfer rather than cross-size transfer.

Our evaluations focus on classification-centered Attention Transfer settings, strictly following the protocol of [8], including ImageNet-1K, iNaturalist, and ImageNet-based out-of-distribution benchmarks. These settings test whether the observed failure boundary persists across different training durations, transfer datasets, and distribution shifts. Applications where attention plays a different operational role, such as dense prediction, vision-language transfer, or cross-resolution deployment, may require task-specific validation before applying the same conclusion. Our mechanistic evidence is empirical: component-level decomposition, native-architecture rescue, native-scratch controls, and teacher-student attention divergence consistently support architectural compatibility as the primary mechanism behind the observed failure. A complete formal theory of Attention Transfer remains an important direction for future work.

D Broader Impacts

This work is primarily a methodological study of when Attention Transfer between pre-trained ViTs succeeds or fails. Its direct positive impact is to encourage more careful reporting of teacher-student architectural assumptions in transfer and distillation studies, and to help the community avoid ineffective transfer attempts under mismatched architectures. A potential negative impact is misinterpretation: our findings should not be read as "Attention Transfer does not work", but rather that Attention Transfer is effective only when the student architecture is compatible with the teacher. Finally, this paper relies on existing public pre-trained weights, some of which are trained on large-scale curated or web-scale datasets and may inherit dataset biases. Architectural compatibility can make transfer more effective, but it does not remove biases or safety concerns inherited from the original teacher weights.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contribution and scope. We revisit the recent claim that Attention Transfer is sufficient to recover the full benefit of the pre-trained weights of Vision Transformers and reveal that Attention Transfer is not universally effective. Further experiments confirm that our findings refine the prevailing understanding of attention in ViT representations: attention is sufficient only when the student architecture matches the teacher.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of our work in Sec. C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: We did not include any theoretical results in this work.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have included the experimental setups in Sec. 2.2 and Sec. A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To ensure reproducibility, we will open-source the code upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included the implementation details in Sec. 2.2 and Sec. A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included the statistical significance analysis of our main results in Sec. B.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included sufficient information on the computing resources needed to reproduce the experiments in Sec. A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research adheres to all ethical guidelines required by NeurIPS.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the broader impacts of our work in Sec. D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks to the best of our knowledge.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and pre-trained weights used in our work are publicly available, and we have cited the respective literature for each of them. Any researcher can download these datasets and pre-trained weights from the provided sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets. We will make the code public after the paper is accepted.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.